A Structured Speech Model Parameterized by Recursive Dynamics and Neural Networks

Roberto Togneri[1], *Li Deng*[2]

[1]School of Electrical, Electronic and Computer Engineering, The University of Western Australia
[2]Microsoft Research, One Microsoft Way, Redmond WA 98052, USA
roberto@ee.uwa.edu.au, deng@microsoft.com

Abstract

We present in this paper an overview of the Hidden Dynamic Model (HDM) paradigm, exemplifying parametric construction of structure-based speech models that can be used for recognition purposes. We explore a general class of the HDM that uses recursive, autoregression functions to represent the hidden speech dynamics, and uses neural networks to represent the functional relationship between the hidden and observed speech vectors. This type of state-space formulation of the HDM is reviewed in terms of model construction, a parameter estimation technique, and a decoding method. We also present some typical experimental results on the use of this type of HDMs for phonetic recognition and for automatic vocal tract resonance tracking. We further provide analyses on the computational complexity (for decoding) and the parameter size of the HDM in comparison with the HMM. Finally, we discuss several key issues related to future exploration of the HDM paradigm.

Index Terms: hidden dynamic model, recursive form of dynamics, neural network, nonlinear mapping, formant tracking, phonetic recognition

1. Introduction

Hidden Dynamic Model (HDM) is one major type of structure-based statistical models designed for speech recognition, where recursive forms of time-varying functions are used to parameterize the un-observed (i.e., hidden) speech dynamics and consequently the observed acoustic feature sequences [2, 11, 12, 13, 14]. The HDM attempts to represent the intrinsic dynamics in the human speech production system in an effort to address some of the known weaknesses of the current hidden Markov modeling (HMM) paradigm. Such weaknesses include the HMM's inability to adequately model long-span coarticulation and phonological changes without resorting to a large number of unstructured, context-dependent parameters. With the current HMM paradigm this can only be achieved by using copious amounts of training data and sophisticated clustering algorithms to reliably estimate the many parameters.

The HDM encompasses a family of related modeling paradigms, which have in common the adoption of a more structured characterization of the underlying speech production dynamics. The use of such a structured model results in far fewer parameters that need to be estimated. With a properly constructed model structure, these models have the potential of being applied to more difficult speech recognition tasks: larger vocabularies, less constrained task grammars, larger populations of speakers, and different speaking styles. However the key to success is a model structure which correctly reflects both the observed dynamics of speech and the underlying articulatory constraints, parameters which can be uniquely determined and reliably estimated, and efficient algorithms for recognition or scoring.

The organisation of this paper is as follows. In Section 2 we briefly review the different HDM modeling types from the literature. In Section 3 we present the model formulation, and training and decoding algorithms for the recursive form of the HDM. In Section 4 we present some typical results from this model. In Section 5 we discuss issues and limitations of the current formulation and suggested future investigations. We conclude the paper in Section 6.

2. Overview of HDM Types

All structured dynamic models attempt to capture the long-span contextual properties of speech by imposing continuity constraints on hidden dynamic quantities which can be mapped back to the observed acoustic features. That is, rather than impose constraints directly on the high-dimensional acoustic feature data, physiologically or phonetically motivated features are used for constraints correlating closely with the underlying speech production mechanisms from which such constraints arise naturally. One of these features are articulatory vectors directly related to speech production [1, 9]. Although these features are ideal, their reliable estimation requires the availability of X-ray or MRI data of the speech production articulators for the complex mapping to the acoustic features and this has greatly limited their use. An alternative feature set which has enjoyed more widespread use has been the vocal-tract resonances (VTRs) observed usually as the formants in voiced sounds.

How the VTR dynamics are represented constitute the main differences between the different modeling paradigms. All models assume that an utterance comprises a sequence of regions or segments which are characterised and "controlled" by VTR "target" values. In non-recursive implementations, the VTR dynamic is derived by noncausal filtering or smoothing of a sequence of constant target values to yield a dynamic trajectory which includes some form of co-articulatory smoothing. The ensuing VTR dynamic is then mapped to the observable features either by a nonlinear mapping function (e.g., [12]), or by an analytical function [4, 5]. Alternatively, in recursive implementations, the VTR dynamic is modeled by a target-directed recursive continuous-valued "state" equation. This can be formulated in a state-space form allowing standard algorithms to

be used for the parameter estimation. The observation equation describes the mapping from the VTR dynamics to the observable features, including linear mappings and mixture of linear mappings [6, 11] and nonlinear mappings [2, 13]. In the following sections we discuss the recursive, state-space HDM with nonlinear mapping which constitutes the most general form of this paradigm.

3. State-Space HDM with Neural-Net Mapping

In the state-space model with recursively defined hidden dynamics, a causal and linear first-order "state" equation is typically used to describe the VTR dynamics according to

$$\mathbf{z}(k+1) = \mathbf{\Phi}^j \mathbf{z}(k) + (\mathbf{I} - \mathbf{\Phi}^j)\mathbf{t}^j + \mathbf{w}(k), \quad j = 1, 2, \ldots, J_P \tag{1}$$

where $\mathbf{z}(k)$ is the low-dimensional "state" vector at discrete time step k, $\mathbf{\Phi}^j$ and $\mathbf{t}^j$ are the system matrix and target vector associated with phone regime j. Both $\mathbf{\Phi}^j$ and $\mathbf{t}^j$ are a function of time k via their dependence on j. The $\mathbf{w}(k)$ is the discrete-time state noise, modeled by an IID, zero-mean, Gaussian process with covariance matrix $\mathbf{Q}$. The observation equation in the model is nonlinear, noisy, and static, and is described by

$$\mathbf{o}(k) = h^{(r)}[\mathbf{z}(k)] + \mathbf{v}(k) \tag{2}$$

where the acoustic observation $\mathbf{o}(k)$ consists of Mel-Cepstra or MFCC (Mel-Frequency Cepstral Coefficients) measurements, and $\mathbf{v}(k)$ is the additive observation noise modeled by an IID, zero-mean, Gaussian process with covariance matrix $\mathbf{R}$. The multivariate nonlinear mapping, $h^{(r)}[\mathbf{z}(k)]$, is implemented by multiple switching MLP (Multi-Layer Perceptron) neural networks, with each MLP associated with a distinct manner (r) of articulation of a phone.

A version of the (generalized) EM algorithm requiring an EKF smoother for the E-step and derivation of estimates for the M-step has been derived and analysed elsewhere and the reader is referred to the relevant literature [2]. An alternative formulation proposed in [14] uses available VTR data (e.g. VTR measurements derived from formant tracker software) to independently estimate the parameters of the state equation, $\Theta_s = \{\mathbf{\Phi}^j, \mathbf{t}^j, \mathbf{Q}\}$, and observation equation, $\Theta_o = \{W_{om}, w_{mi}, \mathbf{R}\}$, where $\{W_{om}, w_{mi}\}$ are the MLP neural network weights. For each phone regime j of interest we obtain the VTR measurements, denoted by $\bar{\mathbf{z}}(k)$, from the phonetically transcribed utterance segments of total length N frames (i.e. $k = 0, 1, ..., N$) for that phone, to yield the sufficient statistics:

$$\mathbf{A} \approx \sum_{k=0}^{N-1} \bar{\mathbf{z}}(k), \qquad \mathbf{B} \approx \sum_{k=0}^{N-1} \bar{\mathbf{z}}(k+1),$$
$$\mathbf{C} \approx \sum_{k=0}^{N-1} \bar{\mathbf{z}}(k)\bar{\mathbf{z}}(k)', \; \mathbf{D} \approx \sum_{k=0}^{N-1} \bar{\mathbf{z}}(k+1)\bar{\mathbf{z}}(k)' \tag{3}$$

$$\mathbf{G} \approx \sum_{k=0}^{N-1} \bar{\mathbf{z}}(k+1)\bar{\mathbf{z}}(k+1)'. \tag{4}$$

From [14] we have the following in the M-step estimation:

$$\widehat{\mathbf{\Phi}} = \mathbf{X}\mathbf{Y}^{-1}, \quad \widehat{\mathbf{t}} = \frac{1}{N}(\mathbf{I} - \widehat{\mathbf{\Phi}})^{-1}(\mathbf{B} - \widehat{\mathbf{\Phi}}\mathbf{A}) \tag{5}$$

where $\widehat{\mathbf{\Phi}}$ is the estimate of the system matrix, $\widehat{\mathbf{t}}$ is the estimate of the target vector, and:

$$\mathbf{X} = \mathbf{B}\mathbf{A}^{'} - N\mathbf{D}, \quad \mathbf{Y} = \mathbf{A}\mathbf{A}^{'} - N\mathbf{C}$$

Furthermore we also form estimates for the noise covariances:

$$\widehat{\mathbf{Q}} = \frac{1}{N}\sum_{k=0}^{N-1} E[\mathbf{e}_{k1}\mathbf{e}'_{k1}|\mathbf{o}, \Theta], \;\; \widehat{\mathbf{R}} = \frac{1}{N}\sum_{k=0}^{N-1} E[\mathbf{e}_{k2}\mathbf{e}'_{k2}|\mathbf{o}, \Theta] \tag{6}$$

where it can be shown that:

$$E\left[\mathbf{e}_{k1}\mathbf{e}_{k1}^{'}|\mathbf{o}, \mathbf{\Theta}\right] = \mathbf{G} - \mathbf{D}\mathbf{\Phi}^{'} - \mathbf{B}\mathbf{t}^{'} - \mathbf{\Phi}\mathbf{D}^{'} + \mathbf{\Phi}\mathbf{C}\mathbf{\Phi}^{'} + \mathbf{\Phi}\mathbf{A}\mathbf{t}^{'} - \mathbf{t}\mathbf{B}^{'} + \mathbf{t}\mathbf{A}^{'}\mathbf{\Phi}^{'} + \mathbf{N}\mathbf{t}\mathbf{t}^{'} \tag{7}$$

$$E\left[\mathbf{e}_{k2}\mathbf{e}_{k2}^{'}|\mathbf{o}, \mathbf{\Theta}\right] = [\mathbf{o}(k) - h(\overline{\mathbf{z}}(k)]\,[\mathbf{o}(k) - h(\overline{\mathbf{z}}(k)]^{'} \tag{8}$$

Finally the MLP weights, $\{W_{om}, w_{mi}\}$, are trained using a standard backpropagation algorithm given the MFCC observations as the desired output sequence, $\mathbf{o}(k)$, and the VTR measurements, $\bar{\mathbf{z}}(k)$, as the corresponding input sequence.

An important quantity that needs to be calculated from the state-space model formulation is the likelihood of the observation sequence given the parameters, $L(\mathbf{o}|\Theta)$. Calculation of $L(\mathbf{o}|\Theta)$ is based on using a single Gaussian to approximate the distribution of the output of the nonlinear dynamic system. This results in an expression based on the pseudo-innovation sequence and its covariance.

4. Experimental Results

While full decoding algorithms using the HDM outlined in Section 3 have not yet been feasible, the HDM can be used to rescore transcriptions, especially in terms of N-best list transcriptions which can be provided by an offline HMM. Results using N-best rescoring on the phone recognition task have been reported for earlier implementations of the HDM [2, 13], including a variant of the HDM using mixture of linear mappings in place of the nonlinear, neural network mapping [11]. The results of a similar N-best rescoring are summarised by Table 1 for the HDM described here trained on the complete TIMIT training data (4620 utterances spoken by 326 male and 136 female speakers) and evaluated on all 1620 utterances (112 males and 56 females) from the TIMIT test data. For the HMM, observations consisted of 39 dimensional static, delta and delta-delta MFCC features which were used to train 3-state, 5-Gaussians/state, triphone models. For the HDM, observations consisted of 13 dimensional MFCC static features and 3-dimensional hidden states (corresponding to the first three VTR components) requiring, for each phone, a 3-input, 12-hidden, and 12-output MLP neural network, a 3-dimensional target vector, and a 3-dimensional system matrix which was assumed diagonal.

Table 1 summarizes a set of results we recently obtained on the standard TIMIT phonetic recognition task, comparing HDM with HMM systems. The results show that HDM is comparable to the HMM in cases where the reference transcription is not included in the 100-best list. However when the reference is included there is a reduction of around 17% in the Word Error Rate (WER) compared to the HMM. This is consistent with previously reported findings for other tasks. When the HDM

	100-best	100-best+ref
Oracle	22.9 (*97.9*)	0.0 (*0.0*)
HMM	31.8 (*100.0*)	30.9 (*97.3*)
HDM	31.4 (*99.9*)	25.0 (*81.9*)

Table 1: Performance for 100-best rescoring measured by Word Error Rate and Sentence Error Rate (italics) or the standard TIMIT phone recognition task using HDM vs. HMM.

is exposed to the correct reference transcription the continuity condition imposed on the VTR state implies a relatively high likelihood score when the transcription is correct, and a significantly improved performance in comparison to an HMM. However when the correct reference is not available any one substitution, insertion or deletion error will propagate to subsequent segments due to the continuity constraint, resulting in a much lower likelihood and reduced ability of the HDM to discriminate between transcriptions with only a few errors and those with many more errors.

As the HDM has been predicated on its ability to model the underlying production model through the hidden dynamics, the work reported in [14] compared the VTR dynamics generated by the HDM with the formant tracks generated from a standard formant tracker software (wavesurfer). The spectrogram plot from Figure 1 is that for one typical utterance from the TIMIT data superimposed by formant tracks and the estimated VTR sequences by the HDM. First, the HDM-VTR sequence (Figure 1(b)) closely follow the formant tracks (Figure 1(a)). Second, in unvoiced regions where the formant tracker fails and produces noisy tracks, the HDM VTR sequence is smoother due to the inherent constraint on the VTR dynamics as a consequence of Eqn. 1. This can be demonstrated by considering only the effect of the predictor step in the EKF recursion (Figure 1(c)) which effectively implements the dynamics imposed by Eqn. 1 without any correction due to the observations. By comparing Figure 1(c) with Figure 1(a) it is also evident that the first-order, target-directed state equation is a reasonable model for the VTR dynamics given its close correspondence to the respective formant tracks.

An important characteristic of the HDM which arises from the structured modeling approach is the reduced number of parameters that need to be estimated compared with an HMM. For the HMM used in the phonetic recognition experiments based on 3-state, 5-Gaussian/state, tri-phone models, 42 distinct phone models, and 39-dimensional feature vectors subject to state tying and regression tree clustering, the number of parameters that need to estimated are given by:

State transitions = 42 x 6 = 252
Mixture weights = 9725
Means and variances= 9728 x 39 x 2 = 758,784
TOTAL PARAMETERS = 768,761

For the HDM based on 42 distinct phone models, 3-dimensional target and system matrix values (assuming a diagonal system matrix), and 3x12x13 MLPs per phone model, the following number of parameters need to be estimated:

Target and System matrix values = 42 x 3 x 2 = 252
MLP weights = 42 x (12x4 + 13x13) = 9114
TOTAL PARAMETERS = 9366

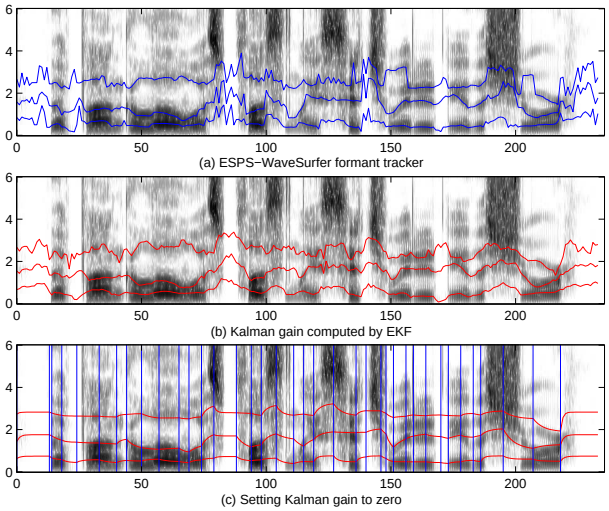

Figure 1: Spectrogram of a TIMIT utterance superimposed with (a) formant tracker data, and recovered VTR sequences from the full HDM (b) and from from the "partial" HDM using only the predictor step of the EKF recursion (c).

These results are summarised in Table 2. A remarkable attribute of the HDM is the number of parameters that need to be estimated which are only 1.2% of the total number of parameters required to estimate a comparable HMM.

Another consequence of the structured modeling paradigm for the HDM is the requirement for more sophisticated training and decoding algorithms. With N-best rescoring, the HDM requires one iteration of the EKF recursion, including calculation of the Jacobian and inverse covariance matrices, whereas the HMM requires one iteration of the Viterbi algorithm where the model sequence is known (only the state path is unknown). An estimate of the required number of multiplications can be approximated for both the HDM and HMM and these results are also summarised in Table 2. Not surprisingly the computational complexity of the HMM is a tenth of that required for the HDM. With more efficient matrix multiplication and matrix inversion algorithms, the computational complexity of the HDM can be improved but the EKF recursion will still be a heavy computational burden compared with the Viterbi algorithm.

	HMM	HDM
Number of Parameters	768761	9366
Decoding Complexity	585	5897

Table 2: Comparison of the number of parameters and computational complexity (for decoding) between HMM and HDM. The computational complexity is measured by the total number of required multiplications for a fixed-length utterance.

5. Issues for Further Exploration

Although the HDM formulation represents a more structured modeling paradigm with a much reduced number of parameters compared to the HMM, there are several issues that need to be further investigated to improve the efficiency and applicability of the model.

The first issue to consider is whether a general nonlinear mapping (such as that represented by a neural network) between the hidden VTR values and the MFCC observation features is necessary. As indicated in [11] a mixture of linear models may be able to provide a similar performance. More investigation is needed to determine the most appropriate forms of the functional mapping that faithfully represents the "physical" relationship between the hidden and observed vectors in the HDM. The mapping functions explored in the past include a MLP or RBF neural network [2, 7, 11, 12], a simple linear mapping [6], a mixture of linear mappings [11], a fixed and parameter-free nonlinear function [3, 4, 5], and codebook mapping constructed from actual articulatory and acoustic data pairs [1].

If one does assume a nonlinear mapping for the observation process then the EKF recursion may not be the best state estimation algorithm to choose. Not only is the EKF computationally expensive but it is only accurate to the first-order due to the linearisation of the Taylor series expansion of the nonlinearity. The Unscented Kalman Filter (UKF) [8], on the other hand, is up to second-order or even third-order accurate in the nonlinearity without any additional computations. Furthermore with the UKF there is no need to derive an expression for the Jacobian, allowing more complex nonlinearities to be investigated.

A more serious problem with the HDM is the fact that an efficient decoding algorithm cannot be easily applied due to the continuity constraint. Alternative formulations like the path-stack algorithm have been proposed in [10] and although much less efficient than the Viterbi algorithm and not optimal, the path-stack algorithm should be investigated further. Alternatively, lattice search rescoring as presented in [4] can at least be used to provide a richer set of transcriptions than N-best rescoring. In the final analysis, to be competitive with or superior to an HMM, the HDM will need to efficiently decode an unknown utterance without any prior knowledge of possible phone segmentations.

6. Discussion and Conclusion

The HMM has been the dominant technology for acoustic modeling in speech recognition, but its weaknesses arising from a number of its inherent assumptions impede the achievement of high performance. One prominent weakness in current HMMs is the handicap in representing long-span temporal dependency in the acoustic feature sequence of speech, reflecting speech coarticulation and reduction. This inadequacy is explicitly addressed by the HDM, a structure-based parametric model for speech dynamics, where several different implementations have appeared in the literature in the past. This paper overviews one principal implementation type of the HDM idea, where the hidden dynamics are represented by recursive-form, autoregressive-style, and target-directed temporal functions, and the relationship between the hidden dynamic vectors and the observed acoustic vectors are represented by a neural network. This implementation is quite different from that of hidden trajectory models [4, 5], where the hidden dynamics are represented by a non-recursive form of the "filtering" function, and the mapping from the hidden to the observed vectors is accomplished by a fixed nonlinear function exploiting the "physical" relationship between VTRs and cepstra.

Future advancement in HDM development will likely come from careful exploration of the several issues discussed in Section 5, and from integration of other types of long-span modeling that can effectively incorporate many other sources of phonetic knowledge in addition to coarticulation modeling as has been the focus in the HDM development to date.

7. References

[1] C.S. Blackburn, S. Young, "Enhanced speech recognition using an articulatory production model trained on X-ray data", Computer Speech and Language, Vol. 15, 2001, pp. 195-215

[2] L. Deng, J. Ma, "Spontaneous speech recognition using a statistical coarticulatory model for the vocal-tract-resonance dynamics", J. Acoust. Soc. Am., Vol. 108, No. 6, 2000, pp. 3036-3048.

[3] L. Deng, X. Li, D. Yu, A. Acero, "A hidden trajectory model with bi-directional target-filtering: cascaded vs. integrated implementation for phonetic recognition", Proc. ICASSP, March 2005, pp. 337-340.

[4] L. Deng, D. Yu, X. Li, A. Acero, "A long-contextual-span model of resonance dynamics for speech recognition: parameter learning and recognizer evaluation", IEEE Workshop on ASRU, Dec. 2005, pp. 145-150.

[5] L. Deng, D. Yu, A. Acero, "Structured speech modeling", IEEE Trans. Audio, Speech and Language Processing, Vol. 14, No. 5, 2006, pp. 1492-1504.

[6] J. Frankel, S. King, "Speech recognition using linear dynamic models", IEEE Trans. Audio, Speech and Language Processing, Vol. 15, No. 1, 2007, pp. 246-256.

[7] P.J.B. Jackson, B.H. Lo, M.J. Russell, "Data-driven, nonlinear, formant-to-acoustic mapping for ASR", Electronics Letters, Vol. 38, No. 13, 2002, pp. 667-669

[8] S.J. Julier, J.K. Uhlmann, "Unscented filtering and nonlinear estimation", Proc. of the IEEE, Vol. 92, No. 3, 2004, pp. 401-422.

[9] L.J. Lee, P. Fieguth, L. Deng, "A functional articulatory dynamic model for speech production", Proc. ICASSP, Mar. 1999, pp. 797-800.

[10] J.Z. Ma, L. Deng, "A path-stack algorithm for optimizing dynamic regimes in a statistical hidden dynamic model of speech", Computer Speech and Language, Vol. 14, 2000, pp.101-114.

[11] J.Z. Ma, L. Deng, "Target-directed mixture dynamic models for spontaneous speech recognition", IEEE Trans. Speech and Audio Processing, Vol. 12, No. 1, 2004, pp. 47-58.

[12] H.B. Richards, J.S. Bridle, "The HDM: A segmental hidden dynamic model of coarticulation", Proc. ICASSP, Mar. 1999, pp. 357-360.

[13] R. Togneri, L. Deng, "An EKF-based algorithm for learning statistical hidden dynamic model parameters for phonetic recognition", Proc. ICASSP, May 2001, pp. 465-468.

[14] R. Togneri, L. Deng, "A state-space model with neural-network prediction for recovering vocal tract resonances in fluent speech from Mel-cepstral coefficients", Speech Communication, Vo. 48, 2006, pp.971-988.